

1 Probability Theory and Statistics

1.1 Probability spaces and random variables

1.1.1 Introduction

- **Definition (Probability space)**

The triple $(\Omega, \mathcal{F}, P)$ where

- Ω is a set of *all possible outcomes*, ω of some experiment (observation), e.g. tosses of a coin;
- $\mathcal{F}$ is a σ -algebra, i.e., a *collection of subsets of Ω* such that:
 - (i) $\Omega \in \mathcal{F}$
 - (ii) if $A \in \mathcal{F}$ then $A^C \in \mathcal{F}$
 - (iii) if $A_1, A_2, \dots \in \mathcal{F}$ then $\cup_i A_i \in \mathcal{F}$
- P is a *probability measure* (see below).

We call the subsets of Ω events. The largest σ -algebra on Ω is the *set of all subsets of Ω* , the smallest one is $\{\emptyset, \Omega\}$. Think why! Ok, good to have a space but we need to somehow quantify the ‘size’ of subsets in it, i.e. we need some ‘measure’. This is given by the function P defined below:

- **Definition (Probability measure)**

A probability measure is a real-valued function P defined over $\mathcal{F}$ and such that:

- (i) for any $A \in \mathcal{F}$, $P(A) \geq 0$
- (ii) $P(\Omega) = 1$, $P(\emptyset) = 0$.
- (iii) (countable additivity) If $\{A_j\}$ is a countable collection of *disjoint* sets in $\mathcal{F}$, i.e., $A_{j1} \cap A_{j2} = \emptyset$, then $P(\cup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} P(A_j)$.

If A is an event (particular $\omega(s)$ being realized), then $P(A)$ is called the **probability** of event A .

- **Definition (Support)**

A *support* of P is any set $A \in \mathcal{F}$ s.t. $P(A) = 1$.

1.1.2 Properties of probability measures

- **1. Monotonicity:** Let P be a probability measure and $A, B \in \mathcal{F}$. If $A \subset B$ then $P(A) \leq P(B)$.
- **2. Inclusion/exclusion:**

$$P(\cup_{k=1}^n A_k) = \sum_{i=1}^n P(A_i) - \sum_{i < j} P(A_i \cap A_j) + \sum_{i < j < k} P(A_i \cap A_j \cap A_k) + \dots (-1)^{n+1} P(A_1 \cap A_2 \dots \cap A_n)$$

This property implies also that $P(A^C) = 1 - P(A)$.

- **3. Continuity:** If $\{A_n\} \in \mathcal{F}$ and $A \in \mathcal{F}$ and $A_n \uparrow A$ then $P(A_n) \uparrow P(A)$. (this means the sequence of sets A_n , subsets of A converges to the set A). The same is true for convergence from above.
- **4. Countable sub-additivity:** If $A_1, A_2, \dots$ and $\cup_{k=1}^{\infty} A_k \in \mathcal{F}$ then $P(\cup A_k) \leq \sum P(A_k)$ (note: here A_k are not necessarily disjoint).

1.1.3 Random variables and distributions

- **Definition (Random Variable)**

A *random variable* (RV) on a probability space $(\Omega, \mathcal{F}, P)$ is a real-valued function $X = X(\omega)$, $\omega \in \Omega$ such that for any $x \in \mathbb{R}$, the set $\{\omega : X(\omega) < x\} \in \mathcal{F}$. When a function satisfies the latter condition we say it is $\mathcal{F}$ -measurable.

A **random vector** is simply a vector of random variables (i.e., vector of functions). Notice that random variables assign real numbers to outcomes. We wish to determine the probability that a real valued function is less than some number, i.e., we are interested in the *probability measures* of sets like $\{\omega : X(\omega) < x\}$.

- **Example**

Consider throwing a die. The event space is $\Omega = \{1, 2, \dots, 6\}$. A possible σ -algebra is $\mathcal{F} = [\{1, \dots, 6\}, \emptyset, \{1, \dots, 4\}, \{5, 6\}]$. A possible probability measure is to assign $P = 1/6$ to each possible outcome ω_i , $i = 1, \dots, 6$ and extend it (using the rules above) to all sets in $\mathcal{F}$. Notice that the probability measure needs to be defined on all sets in $\mathcal{F}$! Now, consider the function:

$$X(\omega) = \begin{cases} 0 & \text{if } \omega \in \{1, 2, 3\} \\ 1 & \text{otherwise} \end{cases}$$

Is $X(\omega)$ a random variable? (Verify this as an exercise using the above definition). What if we had $\omega \in \{1, 2, 3, 4\}$ instead in the first row above?

- **Borel σ -algebra**

Let $\Omega = \mathbb{R}$ and assume we are interested in defining probabilities over open intervals in $\mathbb{R}$, including $(-\infty, \infty)$. It turns out that we can construct a σ -algebra of all open intervals on the real line, which we call the *Borel σ -algebra* on $\mathbb{R}$ and denote by $\mathbb{B}^1$. Note that by the definition of σ -algebra, $\mathbb{B}^1$ must contain also all closed, all semi-open intervals and singletons on $\mathbb{R}$. (think why!)

Next, we proceed by characterizing random variables in more detail.

- **Definition (Distribution function)**

The *cumulative distribution function (cdf)*, $F(x)$ of the random variable $X(\omega)$ is defined as:

$$F(x) = P(\{\omega : X(\omega) \leq x\}) \text{ for } -\infty < x < \infty$$

We often write as a shorthand $P(X \leq x)$ instead of $P(\{\omega : X(\omega) \leq x\})$.

- **Properties of the cdf**

1. $F(-\infty) = 0$ and $F(\infty) = 1$.
2. F is non-decreasing and continuous from the right.
3. $F(x^-) \equiv \lim_{t \uparrow x} F(t)$ and $F(x^+) \equiv \lim_{t \downarrow x} F(t)$ exist and are finite.
4. F is continuous at x iff $F(x^-) = F(x^+) = F(x)$.
5. The only possible discontinuities in F are jumps up.
6. The set of discontinuity points of F is countable.

- **Probability and the cdf**

We have the following relationship between the probability measure P and the cdf F

1. $P(X > x) = 1 - F(x)$
2. $P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$
3. $P(X < x) = F(x^-)$
4. $P(X = x) = F(x^+) - F(x^-)$ (it is positive only if there is a jump at x).

1.1.4 Types of random variables

- **Definition (discrete random variable)**

A discrete RV, X is a RV which can take *only finitely or countably infinitely many values* $x_1, x_2, \dots$ with probabilities $p_1, p_2, \dots, p_j > 0$ and $\sum p_j = 1$. Its cdf is called a discrete distribution function and is a step function with jumps at $x_1, x_2, \dots$. We can also define the *probability function* for X as $f(x) = P(X = x)$. Clearly, $f(x_i) \geq 0$ and $\sum f(x_i) = 1$.

- **Definition (continuous random variable)**

A continuous RV is one that has a *continuous cdf*. Then there exists a function f s.t. $\forall x, y$ with $x < y$ $F(y) - F(x) = \int_x^y f(t)dt$ and F has a derivative equal to f almost everywhere (except on a set of measure 0). The function f is called the *probability density function (pdf)* of X and, as we see, it is only defined up to a set of measure zero, i.e., it is not unique. Unlike the discrete RVs there is no direct connection between $f(x)$ and $P(X = x)$ as the latter is always zero for continuous RVs.

Finally, a *mixed random variable* is for example one that has a continuous part and mass at some point.

• Functions of random variables

Let X is a RV and let us look at the random variable $Y = g(X)$. Suppose we are interested in the distribution of Y , i.e., want to find $f(y)$.

Case 1: X is discrete with pf $f_x(x)$. We have:

$$f_y(y) = P(g(x) = y) = \sum_{x:g(x)=y} f_x(x)$$

Case 2: X is continuous RV with pdf $f_x(x)$ and $P(a < x < b) = 1$. Let g be a continuous and strictly monotonic function and let $a < x < b$ iff $\alpha < y < \beta$. Then the pdf of $Y = g(X)$, $f_y(y)$ is:

$$f_y(y) = \begin{cases} f_x(g^{-1}(y)) \left| \frac{\partial}{\partial y} g^{-1}(y) \right| & \text{for } \alpha < y < \beta \\ 0 & \text{otherwise} \end{cases}$$

1.1.5 Random vectors

In this section we turn to studying vectors of random variables and their distribution functions. Some new concepts need to be defined in this context.

• Definition (joint distribution function)

Let (X, Y) be a random vector. The *joint distribution function (jdf)* $F(x, y)$ of (X, Y) is defined by:

$$F(x, y) = P(\omega : X(\omega) \leq x \text{ and } Y(\omega) \leq y) \quad -\infty < x, y < \infty$$

If X and Y are both discrete RVs then they also have discrete jdf and we can define their joint probability function as $f(x, y) = P(X = x, Y = y)$, with $f(x_i, y_j) \geq 0$ and $\sum_{i,j} f(x_i, y_j) = 1$.

The jdf is just the “overall” distribution function of the vector (X, Y) and is thus analogous to the cdf in the univariate case. If instead we want to look at the components of the random vector *one at a time* we need to define the following:

- **Definition (marginal distribution function)**

Let (X, Y) be a random vector. The *marginal distribution function* of X is:

$$F(x) = \lim_{y \rightarrow \infty} F(x, y) = \lim_{y \rightarrow \infty} P(X \leq x, Y \leq y)$$

The interpretation is that the mdf of X is its distribution function *integrated over all possible values for Y* .

- **Definition (joint and marginal probability density functions)**

(X, Y) have a continuous jdf if there exists a non-negative function $f(x, y)$ called the *joint probability density function* (joint pdf) of (X, Y) such that for any set $A \subset \mathbb{B}^2 : P((X, Y) \in A) = \int_A f(s, t) ds dt$. The joint pdf satisfies the following:

$f(x, y) \geq 0$, $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(s, t) ds dt = 1$, $f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$ whenever the derivative exists.

The *marginal probability density function* of X is $f(x) = \int_{-\infty}^{\infty} f(x, t) dt$. In the discrete case it is called a marginal probability function and is given by $f(x) = P(X = x) = \sum_y f(x, y)$.

- **Exercise:** Suppose that a point (X, Y) is chosen at random from the rectangle $S = \{(x, y) : 0 \leq x \leq 2, 1 \leq y \leq 4\}$. Determine the joint cdf and pdf of X and Y , the marginal cdf and pdf of X and the marginal cdf and pdf of Y .

Answers: $F(x, y) = \frac{x}{2}(\frac{y-1}{3})$ on $[0, 2] \times [1, 4]$ (think how you'd define it outside that area). The joint pdf is then $\frac{\partial^2 F(x, y)}{\partial x \partial y} = \frac{1}{6}$. We also have $F(x) = \frac{x}{2}$ and $F(y) = \frac{y-1}{3}$.

1.1.6 Conditional distributions

The marginal distributions defined above dealt with the *distribution of one of the variables in a random vector for all possible values of the other*. Next we study the distribution of one of the variables given some **fixed value** for the other.

- **Definition (conditional probability function)**

Let X, Y be discrete RVs. The *conditional probability function* of X given $Y = y$ is defined as:

$$f(x|y) = \begin{cases} \frac{f(x, y)}{f(y)} & \text{for } f(y) > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $f(x, y)$ is the joint pf of (X, Y) and $f(y)$ is the marginal pf of Y .

- **Definition (conditional probability density function)**

Let X, Y be continuous RVs. The conditional probability density function of X given $Y = y$ is defined as:

$$f(x|y) = \begin{cases} \frac{f(x, y)}{f(y)} & \text{for } f(y) > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $f(x, y)$ is the joint pdf of (X, Y) and $f(y)$ is the marginal pdf of Y .

- **Example:** suppose we have a fair coin, which we toss twice. Assign 0 if tails occur and 1 if heads. Define the RV X as the *sum of the outcomes* of the two throws and Y as the difference between the first and second throw outcome.

Q1: what are the possible values for X , for Y ?

Q2: what is the joint probability distribution of X and Y .

			X	
		0	1	2
	-1	0	1/4	0
Y	0	1/4	0	1/4
	1	0	1/4	0

Q3: what is the marginal probability function of X ? of Y ?

Q4: what is the conditional probability function of X given $Y = 0$? What is the conditional pf of Y given $X = 0$?

1.1.7 Expectation

- **Definition (Expectation of a continuous random variable)**

The expectation of a continuous RV with pdf $f(x)$ is defined as:

$$E(X) = \int x f(x) dx$$

if the integral exist. The integration is performed on the support of X .

- **Definition (Expectation of a discrete random variable)**

The expectation of a discrete RV with pf $f(x)$ is defined as:

$$E(X) = \sum_i x_i f(x_i)$$

if the sum exist.

Notice that the above imply that it may be possible in some cases the expectation of a RV not to exist. For example take a Cauchy distributed (pdf $f(x) = \frac{1}{\pi(1+x^2)}$) random variable.

We have $\int_{-\infty}^{\infty} xf(x)dx = \frac{1}{\pi} \ln(1+x^2)|_0^{\infty} = \infty$.

- **Example:** compute the expectation of the *uniformly distributed* random variable on $[a, b]$ with $F(x) = \frac{x-a}{b-a}$ for any $x \in [a, b]$.

- **Properties of expectations**

- (1) integrability: $\int xf(x)dx < \infty$, iff $\int |x|f(x)dx < \infty$
- (2) linearity: $E(aX + bY) = aE(X) + bE(Y)$
- (3) If $\exists$ constant a s.t. $P(X \geq a) = 1$ then $E(X) \geq a$
- (4) If $P(X \leq Y) = 1$ then $E(X) \leq E(Y)$.
- (5) If $P(X = Y) = 1$ then $E(X) = E(Y)$
- (6) If $P(X \geq 0) = 1$ and $E(X) = 0$ then $P(X = 0) = 1$.
- (7) If $P(X = 0) = 1$ then $E(X) = 0$.

- **Expectation of a function of RV**

Let X be a RV with pdf $f(x)$. Suppose we want to find $E(Y)$, where $Y = g(X)$ is a random variable which is a function of X . Then the following is true:

$$E(Y) = E(g(X)) = \int g(x)f(x)dx$$

1.1.8 Moments

The expectation of a RV is just one of the many possible characteristics of its distribution. In general, we can define the so-called *moments* of the distribution of a given RV as the expectations of powers of X or $X - E(X)$. In particular, for $r \in \mathbb{N}$, we call $E(X^r)$ the *r-th raw moment* of X and $E((X - E(X))^r)$ the *r-th central moment*. The r-th raw moment exists if $E(|X|^r) < \infty$. Also if $E(|X|^k) < \infty$ for some k , then $E(|X|^j) < \infty$ for any $j < k$.

Let us look at some frequently used moments:

1. Variance (2-nd central moment)

The variance of a RV is given by $V(X) = E((X - E(X))^2) = E(X^2) - (E(X))^2$. We also call $\sqrt{V(X)}$ the *standard deviation* of X . The following are some useful properties of the variance:

- (i) If a is a constant, $V(a) = 0$.
- (ii) $V(aX + b) = a^2V(X)$, for a, b constants.

2. Skewness (3-rd central moment), $E((X - E(X))^3)$

3. Kurtosis (4-th central moment), $E((X - E(X))^4)$

A related concept is the *mean squared error* of a RV defined as:

$$MSE(X) = E(X - c)^2 = V(X) + (c - E(X))^2$$

for some constant c . The interpretation is that it ‘measures’ the average deviation from c . Clearly the MSE is minimized at $c = E(X)$.

1.1.9 Useful Inequalities

In this section we establish some useful connections between some of the moments of random variables and probabilities of certain events.

- **Theorem 31 (Markov's Inequality)**

Let Y be a non-negative RV, i.e. $P(Y < 0) = 0$ and let k be a positive constant. Then:

$$P(Y \leq k) \leq \frac{E(Y)}{k}$$

- **Theorem 32 (Chebyshev's Inequality #1)**

Let X be a RV, c is a constant and d is a positive constant. Then:

$$P(|X - c| \geq d) \leq \frac{E(X - c)^2}{d^2}$$

- **Theorem 33 (Chebyshev's Inequality #2)**

Let X be a RV with expectation $E(X) = \mu$ and variance $V(X) = \sigma^2$ and d is a positive constant. Then:

$$P(|X - \mu| \geq d) \leq \frac{\sigma^2}{d^2}$$

- **Theorem 34 (Jensen's Inequality)**

Let X be a RV and h is a *concave function*. Then:

$$E(h(X)) \leq h(E(X))$$

1.1.10 Covariance

So far we only studied the characteristics of the distribution of a single random variable. Often, however it is interesting to know how two RVs are related. In order to achieve that we can define the **covariance** of two RVs X and Y as:

$$\text{cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

Notice the closeness with the concept of variance (just put $Y = X$ above). The covariance measures how the two RVs are related to each other, notice that if X and Y are such that if X goes up then Y goes up too, they will have a positive covariance, whereas if they are negatively related they will have a negative covariance.

Properties

- (1) $\text{cov}(X, Y) = E(XY) - E(X)E(Y)$
- (2) $\text{cov}(X + Y, Z) = \text{cov}(X, Z) + \text{cov}(Y, Z)$
- (3) $\text{cov}(aX, Y) = a(\text{cov}(X, Y))$ for a constant.
- (4) $\text{cov}(X, a) = 0$ for a constant.
- (5) If $X_1, X_2, \dots, X_n$ are RVs with $V(X_i) < \infty$, all i , then:

$$V\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n V(X_i) + \sum_{i,j=1, i \neq j}^n \text{cov}(X_i, X_j)$$

1.1.11 Correlation

A related concept to that of covariance is the **correlation** of two RVs defined as:

$$\rho(X, Y) \equiv \frac{\text{cov}(X, Y)}{\sqrt{V(X)V(Y)}} \text{ for } V(X), V(Y) < \infty$$

Notice that the correlation is simply the covariance normalized by the product of standard deviations of the two variables. Because of this normalization we can prove the following:

- **Theorem 35**

The correlation between two RVs is always less or equal to 1 in absolute value, i.e.

$$\rho(X, Y) \in [-1, 1].$$

If $\rho(X, Y) = 1$ we say that the RVs are perfectly positively correlated. If $\rho(X, Y) = -1$ we call them perfectly negatively correlated. If $\rho(X, Y) = 0$ we say that the RVs are *uncorrelated*.

- **Exercise:** Suppose that X is uniformly distributed on the interval $[-1, 1]$ and $Y = X^2$. Show that X and Y are uncorrelated.

Solution: It will be enough to show that $\text{cov}(X, Y) = 0$ (show that the variances are finite first). We have $E(X) = 0$, $F(x) = \frac{x+1}{2}$ and $f(x) = 1/2$. What is the expectation of Y ? We have $E(Y) = \int_{-1}^1 x^2 \frac{1}{2} dx = (1/6 + 1/6) = 1/3$. Thus

$$\text{cov}(X, Y) = E[(X - 0)(X^2 - 1/3)] = E(X^3 - \frac{X}{3}) = E(X^3) - 0$$

We also have $E(X^3) = \int_{-1}^1 x^3 \frac{1}{2} dx = (1/8) - (1/8) = 0$. Thus, the two RVs are indeed *uncorrelated* even though one is a function of the other!

1.1.12 Conditional expectation

Often we may be interested in the expectation of some RV X given that some other variable Y takes some fixed value.

- **Definition (conditional expectation)**

(a) The conditional expectation of a discrete RV X given a discrete RV Y , is:

$$E(X|Y) = \sum_x x f(x|y)$$

(b) The conditional expectation of a continuous RV X given a continuous RV Y , is:

$$E(X|Y) = \int x f(x|y) dx$$

Notice that $E(X|Y)$ is a random variable, taking values $E(X|y)$ when $Y = y$. Related to the concept of conditional expectation we have the following important result.

Exercise: compute $E(X)$, $E(Y)$ and the conditional expectations for some fixed X or Y in the coin toss example above.

- **Theorem 36 (the “Law of iterated expectations”)**

Let X, Y be two random variables with finite expectations. Then:

$$E(Y) = E(E(Y|X))$$

The intuition is the following. $E(Y)$ is the expectation about Y without any information about X , whereas $E(Y|X)$ is the expectation about Y knowing X . Thus in general $E(Y|X) \neq E(Y)$ but *on average*, i.e. $E(E(Y|X))$ it must be equal to $E(Y)$ otherwise the latter would be wrong.

- There is also a similar result for the variance (*the variance decomposition*):

$$V(Y) = V(E(Y|X)) + E(V(Y|X))$$

1.1.13 Independence

- **Definition**

The random variables $\{X_j\}_{j=1}^n$ are *independent* if and only if

$$P(\cap_{j=1}^n (X_j \in A_j)) = \prod_{j=1}^n P(X_j \in A_j)$$

for any $A_j \in \mathcal{B}^1$, $j = 1..n$.

The following results are often used to verify the independence of RVs.

- **Theorem 37**

The random variables $X_1, X_2, \dots, X_n$ are independent if and only if for all $x_1, \dots, x_n$:

- (i) $F(x_1, x_2, \dots, x_n) = \prod_{j=1}^n F(x_j)$, where $F(X)$ is the cdf of X .
- (ii) $f(x_1, \dots, x_n) = \prod_{j=1}^n f(x_j)$, if the marginal pdf's exist.

- **Theorem 38 (Independence of functions of random variables)**

Let $\{X_j\}_{j=1}^n$ are independent RVs and $\{g_j\}$ is a sequence of functions. Then $\{g_j(X_j)\}_{j=1}^n$ are also independent RVs.

1.1.14 Independence vs. uncorrelatedness

- **Theorem 39**

If $X_1, \dots, X_n$ be n independent RVs defined on $(\Omega, \mathcal{F}, P)$ whose respective raw moments of orders $r_1, \dots, r_n$ exists. Then:

$$E(X_1^{r_1} X_2^{r_2} \dots X_n^{r_n}) = E(X_1^{r_1}) E(X_2^{r_2}) \dots E(X_n^{r_n})$$

- **Corollary**

If the RVs X and Y are independent then they are also uncorrelated.

The above is true since $E(XY) = E(X)E(Y)$ by Theorem 39, thus $\text{cov}(X, Y) = 0$. However, **it is not true that if two variables are uncorrelated, they are also independent.** Take a look at the following counter-example.

- **Example (uncorrelatedness does not imply independence)**

Remember our fair coin, which we toss twice. Assign 0 if tails occur and 1 if heads. Define the RV X as the sum of the outcomes of the two throws and Y as the difference between the first and second outcome. The joint probability distribution of X and Y is:

			X	
		0	1	2
	-1	0	1/4	0
Y	0	1/4	0	1/4
	1	0	1/4	0

We have $E(X) = 1$, $E(Y) = 0$, $E(XY) = 0$, i.e., $\text{cov}(X, Y) = 0$, thus X and Y are uncorrelated (verify those claims). However they are not independent since $f(X = 1, Y = 0) = 0$ but $f(X = 1) = 1/2$ and $f(Y = 0) = 1/2$, i.e., $f(1, 0) \neq f_x(1)f_y(0)$.

NOTE: another example was the $Y = X^2$ above. Show it.

- **Example 2:** Suppose that (X, Y) are uniformly distributed on the square $R = \{(x, y) : -6 \leq x \leq 6, -6 \leq y \leq 6\}$. Show that X and Y are independent and hence uncorrelated.

Solution: we have $F(x, y) = \frac{(x+6)}{12} \frac{(y+6)}{12}$ on $[-6, 6] \times [-6, 6]$ (draw a picture, this is about the area of a rectangle). But it is also true that the marginal distributions $F(x) = \lim_{y \rightarrow 6} F(x, Y) = \frac{(x+6)}{12}$ and similarly for Y .

However, under some special assumptions about X and Y independence and uncorrelatedness may be equivalent.

- **Theorem 40**

If X, Y are jointly *normally distributed*, then X and Y being independent is equivalent to X and Y being uncorrelated.

1.2 Convergence of random variables

We consider sequences of random vectors $\{X_n\}_{n=1}^{\infty}$ and we are interested in the convergence of such sequences. Remember that random variables are functions so this is related to the general theory on convergence of functions. Serving different purposes, the following three possible definitions of convergence of random variables exist. They are **not equivalent** but some of them are implied by the others as we will see later on.

- **Definition (almost sure convergence)**

We say that $\{X_n\}$ converges *almost surely* to X (written as $X_n \rightarrow^{a.s.} X$) if and only if:

$$P(\omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)) = 1$$

The above definition requires that X_n converges to X for all $\omega \in \Omega$ except possibly on a set of measure 0.

- **Definition (convergence in probability)**

We say that $\{X_n\}$ converges *in probability* (written as $X_n \rightarrow^p X$) if and only if:

$$\forall \varepsilon > 0, \lim_{n \rightarrow \infty} P(\omega : \|X_n(\omega) - X(\omega)\| > \varepsilon) = 0$$

where $\|a\|$ denotes the Euclidean norm.

- **Definition (convergence in distribution)**

We say that $\{X_n\}$ converges *in distribution* to X (written as $X_n \rightarrow^d X$) if and only if:

$$\lim_{n \rightarrow \infty} F_n(x) = F(x)$$

where F_n is the joint cdf of X_n and F is the joint cdf of X .

The following theorem gives the relationships between the above convergence concepts.

- **Theorem 42 (Relationship between convergence concepts)**

Let $\{X_n\}_{n=1}^{\infty}$ is a sequence of random variables such that the above convergence concepts are well defined. Then:

(i) the following relationships hold:

$$X_n \rightarrow^{a.s.} X \Rightarrow X_n \rightarrow^p X \Rightarrow X_n \rightarrow^d X$$

(ii) If $X_n \rightarrow^d c$ – a constant, then $X_n \rightarrow^p c$.

Let us analyze some of the above relations. We see that $X_n \rightarrow^d X$ is the weakest convergence concept.

1.2.1 Useful Convergence Theorems

The following theorems are very often used in econometrics with various applications.

- **Theorem 43 (Mann-Wald)**

- (i) Let $g : \mathbb{R}^k \rightarrow \mathbb{R}$ be a continuous function. Then $X_n \rightarrow^d X$ implies $g(X_n) \rightarrow^d g(X)$.
- (ii) Let $g : \mathbb{R}^k \rightarrow \mathbb{R}$ be a continuous function. Then $X_n \rightarrow^p X$ implies $g(X_n) \rightarrow^p g(X)$.

- **Theorem 44 (Slutski)**

Let X_n and Y_n be scalar RVs.

- (a) If $X_n \rightarrow^d X$ and $Y_n \rightarrow^d c$ – constant, then:
 - (i) $(X_n, Y_n) \rightarrow^d (X, c)$
 - (ii) $X_n + Y_n \rightarrow^d X + c$
 - (iii) $X_n Y_n \rightarrow^d Xc$
 - (iv) $\frac{X_n}{Y_n} \rightarrow^d \frac{X}{c}$ if $c \neq 0$.

The following result is very useful for finding limiting distributions of functions of random variables.

- **Theorem 45 (Delta method)**

Let $\alpha_n(X_n - b) \rightarrow^d X$ where $\alpha_n \rightarrow \infty$. Let g be a function differentiable at b . Then

$$\alpha_n(g(X_n) - g(b)) \rightarrow^d \frac{\partial g(b)}{\partial b} X$$

Example:

Let $\tilde{\theta}$ is a RV such that $\sqrt{n}(\tilde{\theta} - \theta) \rightarrow^d N(0, \sigma^2)$. Suppose we are interested in the limiting distribution of $\gamma = e^{\theta}$. Let $\tilde{\gamma} = e^{\tilde{\theta}}$. Then by Theorem 45, we have that

$$\sqrt{n}(g(\tilde{\theta}) - g(\theta)) \rightarrow^d g'(\theta)N(0, \sigma^2)$$

$$\text{i.e., } \sqrt{n}(\tilde{\gamma} - \gamma) \rightarrow^d N(0, \gamma^2 \sigma^2)$$

1.3 Laws of large numbers

Let X_i , $i = 1, 2, \dots$ be a collection of random variables with finite expectations given by $\mu_i = E(X_i)$. Let also $\bar{X}_n = \frac{1}{n} \sum_{j=1}^n X_j$. A **Law of large numbers (LLN)** provides conditions on $\{X_i\}_{i=1}^\infty$ such that the derived sequence of partial averages $\{\bar{X}_n\}_{n=1}^\infty$ (notice that this is a sequence of RVs itself) converges in probability (for a Weak LLN) or almost surely (for a Strong LLN) to some limit.

- **Theorem 47 (Khintchin's weak LLN)**

Let X_i , $i = 1, 2, \dots$ are independent and identically distributed (i.i.d) RVs and assume that $E(X_i) = \mu < \infty$. Then:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{p} \mu$$

- **Theorem 48 (Chebyshev's weak LLN)**

Let X_i , $i = 1, 2, \dots$ satisfy $E(X_i) = \mu_i < \infty$, $V(X_i) = \sigma_i^2 < \infty$ and $cov(X_i, X_j) = 0$, all $i \neq j$. If $\lim_{n \rightarrow \infty} V(\frac{1}{n} \sum_{i=1}^n X_i) = \lim_{n \rightarrow \infty} \frac{1}{n^2} \sum \sigma_i^2 < \infty$ then:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{p} E(\bar{X}_n)$$

- **Theorem 50 (Kolmogorov's strong LLN)**

Suppose X_i , $i = 1, 2, \dots$ are independent and with finite variances, given by σ_i^2 . If $\sum_{n=1}^\infty \frac{\sigma_n^2}{n^2} < \infty$ then

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{a.s.} E(\bar{X}_n)$$

Note also that since almost sure convergence implies convergence in probability the strong LLN is also weak.

1.4 Central limit theorems

A Central Limit Theorem (CLT) gives conditions on a sequence of random variables $\{X_i\}_{i=1}^\infty$ such that the derived sequence of partial averages $\{\bar{X}_n\}_{n=1}^\infty$ (where $\bar{X}_n \equiv \frac{1}{n} \sum_{j=1}^n X_j$), when suitably normalized, converges in distribution to a standard normal distribution. As with the LLNs there are many possible CLTs of which the following are the most frequently used.

- **Theorem 50 (Lindberg-Levy CLT)**

Let $X_i, i = 1, 2, \dots$ are *i.i.d.* random variables with $E(X_i) = \mu < \infty$ and $V(X_i) = \sigma^2 < \infty$. Then:

$$\frac{\frac{1}{n} \sum_{i=1}^n X_i - \mu}{\sqrt{\sigma^2/n}} \rightarrow^d N(0, 1)$$

This theorem makes strong assumptions about homogeneity and independence of the RVs. Theorems with weaker homogeneity assumptions but stronger moments assumptions are given as follows.

- **Theorem 51 (Liapounov CLT)**

Suppose $X_i, i = 1, 2, \dots$ are *independent* random variables with $E(X_i) = \mu_i < \infty$ and $V(X_i) = \sigma_i^2 < \infty$ and $\mu_{3i} = E(|X_i - \mu_i|^3) < \infty$. Let $B_n = (\sum_{i=1}^n \mu_{3i})^{1/3}$, $C_n = (\sum_{i=1}^n \sigma_i^2)^{1/2}$. If $\lim_{n \rightarrow \infty} \frac{B_n}{C_n} = 0$ then:

$$\frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{V(\bar{X}_n)}} \rightarrow^d N(0, 1)$$

where $\bar{X}_n \equiv \frac{1}{n} \sum_{j=1}^n X_j$.

- **Theorem 52 (Lindberg-Feller CLT)**

Suppose $X_i, i = 1, 2, \dots$ are *independent* random variables with distribution functions F_i , $E(X_i) = \mu_i < \infty$ and $V(X_i) = \sigma_i^2 < \infty$. Let $C_n = |\sum_{i=1}^n \sigma_i^2|^{1/2}$. If $\lim_{n \rightarrow \infty} \frac{1}{C_n^2} \sum_{i=1}^n \int_{|x - \mu_i| > \varepsilon C_n} (x - \mu_i)^2 dF_i(x) = 0$ for all $\varepsilon > 0$ then

$$\frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{V(\bar{X}_n)}} \rightarrow^d N(0, 1)$$